A COMPARISON OF TWO APPROACHES: MAXIMUM ENTROPY ON THE MEAN (MEM) AND BAYESIAN ESTIMATION (BAYES) FOR INVERSE PROBLEMS

Ali Mohammad-Djafari
Laboratoire des Signaux et Systèmes (CNRS-ESE-UPS)
École Supérieure d'Électricité,
Plateau de Moulon, 91192 Gif-sur-Yvette, France.
E-mail: djafari@lss.supelec.fr

Abstract. To handle with inverse problems, two probabilistic approaches have been proposed: the maximum entropy on the mean (MEM) and the Bayesian estimation (BAYES). The main object of this presentation is to compare these two approaches which are in fact two different inference procedures to define the solution of an inverse problem as the optimizer of a compound criterion.

Key words: Inverse problems, Maximum Entropy on the Mean, Bayesian inference, Convex analysis

1. Introduction

Inverse problems arises in many areas of science and engineering. In fact, rarely, we can measure directly a quantity x and, in general, the unobserved interested x is related to the measured quantity y via a model. In many area this model can be written in the general form $g = \mathcal{A}(x) + n$ or in the discrete case:

$$\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{x}) + \boldsymbol{n}, \tag{1}$$

where $\boldsymbol{y}$ stands for the data, $\boldsymbol{x}$ for the unknown variables and $\boldsymbol{n}$ for the errors (modeling and noise). Since Newton and Gauss, one tries to define a solution to this problem as the optimizer of a criterion, for example the Least Squares (LS):

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \left\{ \|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{x})\|^2 \right\}. \tag{2}$$

But the inverse problems are, in general, ill-posed and the LS criterion may not have a unique optimum or this solution may be very sensitive to noise. Since Tikhonov [**?**], the regularization theory became the main approach to give a satisfactory solution by defining it as the optimizer of a compound criterion:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \{J(\boldsymbol{x})\} \quad \text{with} \quad J(\boldsymbol{x}) = Q(\boldsymbol{x}) + \lambda\Omega(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{x})\|^2 + \lambda\|\boldsymbol{D}\boldsymbol{x}\|^2 \tag{3}$$

or in its more general forms [**?**]:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \{J(\boldsymbol{x})\} \quad \text{with} \quad J(\boldsymbol{x}) = Q(\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{x})) + \lambda\Omega(\boldsymbol{x}, \boldsymbol{m}). \tag{4}$$

The questions then raised on how to choose the functionals Q and Ω and the regularization parameter λ and the default solution $\boldsymbol{m}$.

The probabilistic approaches started to give partial answers to this request. In particular in the Bayesian estimation approach and the maximum a posteriori (MAP) estimate:

$$\widehat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \{p(\boldsymbol{x}|\boldsymbol{y})\} = \arg\min_{\boldsymbol{x}} \{-\log p(\boldsymbol{x}|\boldsymbol{y})\} = \arg\min_{\boldsymbol{x}} \{-\log p(\boldsymbol{y}|\boldsymbol{x}) - \log p(\boldsymbol{x})\}, \tag{5}$$

this choice is: $Q = -\log p(\boldsymbol{y}|\boldsymbol{x})$ and $\lambda\Omega(\boldsymbol{x}) = -\log p(\boldsymbol{x})$. This approach just pushed a little farther the questions which became how to translate our prior knowledge into a probability law and how to determine their parameters. Even, nowadays, there are many tools for the estimation of the hyperparameters [**?**], the main question on how to translate some knowledge about $\boldsymbol{x}$ into a probability distribution stays without a complete answer. The maximum entropy (ME) principle gave partial answers [**?**, **?**, **?**]. See also [**?**] for an extensive discussed bibliography.

At the same time, many authors used the ME principle to find unique solutions to linear inverse problems by considering $\boldsymbol{x}$ as a distribution and the data $\boldsymbol{y}$ as linear constraints on them. Then, assuming that the data constraints are satisfied by a non empty set of solutions, a unique solution is chosen by maximizing the entropy:

$$-\sum_j x_j \log x_j \quad \text{or} \quad -\sum_j x_j \log\left[\frac{x_j}{m_j} - (x_j - m_j)\right], \tag{6}$$

where $\boldsymbol{m}$ is default solution. See for example [**?**, **?**] and the cited references. However, even if in these methods, thanks to convex analysis and Lagrangian techniques, the constrained optimization of 6 can be replaced by an equivalent unconstrained optimization, the obtained solutions satisfy the uniqueness condition of well-posedness but not the stability one [**?**, **?**, **?**].

Recently, some authors [**?**, **?**, **?**, **?**] used the ME principle in a different way by considering $\boldsymbol{x}$ not as a distribution but as the mean value of a random vector $\boldsymbol{X}$ and the data as the constraints on its distribution $\mathrm{d}P(\boldsymbol{x})$. Then, the ME principle is used to define it uniquely and finally the solution $\widehat{\boldsymbol{x}}$ is defined as the expected value of this ME distribution.

Following these authors, some others used, commented and analyzed extensively these ideas [**?**, **?**, **?**, **?**, **?**, **?**], [**?**, **?**, **?**, **?**, **?**] and [**?**, **?**, **?**, **?**]. However, in all these works, the data $\boldsymbol{y}$ were considered as exact constraints and the errors on the data were either neglected or partially token account of. (See however new developments in [**?**].)

More recently, some authors who were more faced with real applications, [**?**, **?**, **?**, **?**] followed the same idea, but by fixing themselves as the objective to use these ideas for describing the solution as the optimizer of a combined convex criteria such as (4) and more on a constructive way to determine these functionals.

The objective of this paper is to make a comparison of the Bayesian approach which we call hereafter BAYES and the maximum entropy in the mean which we refer to as MEM. This comparison is done very pragmatically and is based on the

understanding of the author who does not have pretension to know all the details of the both approaches and will be happy to discuss all the following discussions with the pro of the approaches.

2. Maximum entropy on the mean approach

2.1. Basics

The main references to the basics of this approach are [**?**, **?**]. The original idea and first applications in crystallography are given in [**?**, **?**]. More details and extensions are given in [**?**, **?**, **?**, **?**]. The mathematical aspects of convex analysis and duality theorems are given in [**?**, **?**, **?**, **?**, **?**, **?**].

The following resumes the different steps of the approach:

– Consider a set $\mathcal{C}$, assume that $\boldsymbol{x} \in \mathcal{C}$ and define a reference measure $\mu(\boldsymbol{x})$:

$$\boldsymbol{x} \in \mathcal{C}, \qquad \boldsymbol{m} = \int_{\mathcal{C}} \boldsymbol{x}\, \mathrm{d}\mu(\boldsymbol{x}), \tag{7}$$

where $\boldsymbol{m}$ is the mean value of $\boldsymbol{x}$ under this reference measure.

– Consider $\boldsymbol{x}$ as the mean value of a random vector $\boldsymbol{X}$ for which you assume a probability distribution P:

$$\boldsymbol{x} = \mathrm{E}_P\{\boldsymbol{X}\} = \int_{\mathcal{C}} \boldsymbol{x}\, \mathrm{d}P(\boldsymbol{x}) \tag{8}$$

and the data $\boldsymbol{y}$ as exact equality constraints on it:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}\mathrm{E}_P\{\boldsymbol{X}\} = \int_{\mathcal{C}} \boldsymbol{A}\boldsymbol{x}\, \mathrm{d}P(\boldsymbol{x}). \tag{9}$$

– Determine the distribution P by:

$$\text{maximize} \quad -\int_{\mathcal{C}} \log \frac{\mathrm{d}P(\boldsymbol{x})}{\mathrm{d}\mu(\boldsymbol{x})}\, \mathrm{d}P(\boldsymbol{x}) \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}\mathrm{E}_P\{\boldsymbol{X}\}. \tag{10}$$

The solution is calculated via Lagrangian:

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) &= \int_{\mathcal{C}} \left[\log \frac{\mathrm{d}P(\boldsymbol{x})}{\mathrm{d}\mu(\boldsymbol{x})} - \sum_{i=1}^{M} \lambda_i (y_i - [\boldsymbol{A}\boldsymbol{x}]_i) \right] \mathrm{d}P(\boldsymbol{x}) \\ &= \int_{\mathcal{C}} \left[\log \frac{\mathrm{d}P(\boldsymbol{x})}{\mathrm{d}\mu(\boldsymbol{x})} - \boldsymbol{\lambda}^t (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) \right] \mathrm{d}P(\boldsymbol{x}) \end{aligned} \tag{11}$$

and is given by:

$$\mathrm{d}P(\boldsymbol{x}, \boldsymbol{\lambda}) = \exp\left[\boldsymbol{\lambda}^t [\boldsymbol{A}\boldsymbol{x}] - \log Z(\boldsymbol{\lambda})\right]\, \mathrm{d}\mu(\boldsymbol{x}), \tag{12}$$

where

$$Z(\boldsymbol{\lambda}) = \int_{\mathcal{C}} \exp\left[\boldsymbol{\lambda}^t [\boldsymbol{A}\boldsymbol{x}]\right]\, \mathrm{d}\mu(\boldsymbol{x}). \tag{13}$$

The Lagrange parameters are calculated by searching the unique solution (if exists) of the following system of non linear equations:

$$\frac{\partial \log Z(\boldsymbol{\lambda})}{\partial \lambda_i} = y_i, \quad i = 1, \cdots, M. \tag{14}$$

– The solution to the inverse problem is then defined as the expected value of this distribution:

$$\widehat{\boldsymbol{x}}(\boldsymbol{\lambda}) = \int \boldsymbol{x} \; \mathrm{d}P(\boldsymbol{x}, \boldsymbol{\lambda}). \tag{15}$$

These steps are very formal. In fact, it is possible to determine $\widehat{\boldsymbol{x}}(\boldsymbol{\lambda})$ in a more direct manner. Using the following notations:

$$\boldsymbol{s} = \boldsymbol{A}^t \boldsymbol{\lambda}, \quad G^*(\boldsymbol{s}) = \log Z(\boldsymbol{s}) = \log \int_{\mathcal{C}} \exp\left[\boldsymbol{s}^t \boldsymbol{x}\right] \; \mathrm{d}\mu(\boldsymbol{x}), \tag{16}$$

and

$$H(\boldsymbol{x}) = \max_{\boldsymbol{s}} \{\boldsymbol{s}^t \boldsymbol{x} - G^*(\boldsymbol{s})\}, \quad D(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^t \boldsymbol{y} - G^*(\boldsymbol{A}^t \boldsymbol{\lambda}) \tag{17}$$

it is shown that:

$$\widehat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} \{D(\boldsymbol{\lambda})\} \qquad \text{(Dual criterion)} \tag{18}$$

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x} \in \mathcal{C}} \{H(\boldsymbol{x})\} \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \qquad \text{(Primal criterion)}, \tag{19}$$

$$\widehat{\boldsymbol{x}}(\boldsymbol{s}) = \frac{\mathrm{d}G^*(\boldsymbol{s})}{\mathrm{d}\boldsymbol{s}} \qquad \text{(Explicit relation)}, \tag{20}$$

where:

– Functions G and H depend on the reference measure $\mu(\boldsymbol{x})$;
– $D(\boldsymbol{\lambda})$ is the dual criterion which is function of the data and the function G;
– $H(\boldsymbol{x}) = H(\boldsymbol{x}, \boldsymbol{m})$ is the primal criterion which is a distance measure between $\boldsymbol{x}$ and $\boldsymbol{m}$ which means:
  – $H(\boldsymbol{x}, \boldsymbol{m}) \geq 0, \quad$ and $\quad H(\boldsymbol{x}, \boldsymbol{m}) = 0 \quad$ iff $\quad \boldsymbol{x} = \boldsymbol{m}$;
  – $H(\boldsymbol{x}, \boldsymbol{m})$ is differentiable and convex on $\mathcal{C}$;
  – $H(\boldsymbol{x}, \boldsymbol{m}) = \infty \quad$ if $\boldsymbol{x} \notin \mathcal{C}$.

Now, to be able to go a little more in details, let assume that the reference measure is separable:

$$\mu(\boldsymbol{x}) = \prod_{j=1}^{N} \mu_j(x_j) \tag{21}$$

then, we have:

$$\mathrm{d}P(\boldsymbol{x}, \boldsymbol{\lambda}) = \prod_{j=1}^{N} \mathrm{d}P_j(x_j, \boldsymbol{\lambda}) \tag{22}$$

and

$$G(\boldsymbol{s}) = \sum_j g_j\left(s_j\right), \quad H(\boldsymbol{x}, \boldsymbol{m}) = \sum_j h_j(x_j, m_j), \quad \widehat{x}_j = g_j'(s_j). \tag{23}$$

Replacing $\boldsymbol{s} = \boldsymbol{A}^t\boldsymbol{\lambda}$ we obtain:

$$G(\boldsymbol{\lambda}) = \sum_j g_j\left([\boldsymbol{A}^t\boldsymbol{\lambda}]_j\right), \quad H(\boldsymbol{x},\boldsymbol{m}) = \sum_j h_j(x_j,m_j), \quad \widehat{x}_j = g_j'\left([\boldsymbol{A}^t\widehat{\boldsymbol{\lambda}}]_j\right), \quad (24)$$

where h_j and g_j depend on the reference measure μ_j:

– g_j is the log Laplace transform (Cramer transform) of μ_j:

$$g(s) = \log \int \exp[sx]\ \mathrm{d}\mu(x);$$

– h_j is the convex conjugate of g_j: $\quad h(x) = \max_{\boldsymbol{s}}\{sx - g(s)\}$.

Let give some examples:

	$\mu_j(x)$	$g_j(s)$	$h_j(x,m)$
Gaussian:	$\exp\left[-\frac{1}{2}(x-m)^2\right]$	$\frac{1}{2}(s-m)^2$	$\frac{1}{2}(x-m)^2$
Poisson:	$\frac{m^x}{x!}\exp[-m]$	$\exp[m-s]$	$-x\log\frac{x}{m}+m-x$
Gamma:	$x^{\alpha-1}\exp\left[-\frac{x}{m}\right]$	$\log(s-m)$	$-\log\frac{x}{m}+\frac{x}{m}-1$

When $\mu(\boldsymbol{x})$ is not separable it is very difficult to do the calculation more in details, excepted the Gaussian case $\mu(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{m},\boldsymbol{R}_x)$, where we have:

$$H(\boldsymbol{x},\boldsymbol{m}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^t\boldsymbol{R}_x^{-1}(\boldsymbol{x}-\boldsymbol{m}), \quad G(\boldsymbol{\lambda}) = -\frac{1}{2}\|\boldsymbol{\lambda}\|^2, \quad D(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^t\boldsymbol{y} + \frac{1}{2}\|\boldsymbol{\lambda}\|^2. \quad (25)$$

(See however a new presentation of the method in [**?**] trying to extend the method for taking account of the correlations.)

2.2. Extensions

How to account for the noise: Two approaches have been developed in [**?**, **?**, **?**, **?**]:

– Replacing the exact equality constraints $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ by the following inequalities:

$$|y_i - [\boldsymbol{A}\boldsymbol{x}]_i| < \epsilon, \quad \text{or} \quad \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 < \sigma^2 \quad (26)$$

and using the duality relations they showed:

$$\begin{cases} \widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}}\{H(\boldsymbol{x})\} \text{ s.t. } |y_i - [\boldsymbol{A}\boldsymbol{x}]_i| < \epsilon, \text{ with } H(\boldsymbol{x}) = \sum_j h_j(x_j) \\ \widehat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}}\left\{\tilde{D}(\boldsymbol{\lambda})\right\} \quad \text{with} \quad \tilde{D}(\boldsymbol{\lambda}) = D(\boldsymbol{\lambda}) + \alpha\|\boldsymbol{\lambda}\|^2 \end{cases} \quad (27)$$

where α depends on ϵ or on σ^2.

– Replacing $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ by $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}$ and rewriting it as follows:

$$\boldsymbol{y} = [\boldsymbol{A}|\boldsymbol{I}] \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{n} \end{bmatrix} \longrightarrow \boldsymbol{y} = \tilde{\boldsymbol{A}}\,\tilde{\boldsymbol{x}} \tag{28}$$

and assuming $\mu(\tilde{\boldsymbol{x}}) = \mu_x(\boldsymbol{x})\mu_n(\boldsymbol{n})$ they showed:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathcal{C}} \{\mathcal{Q}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) + \alpha H(\boldsymbol{x})\} \tag{29}$$

$$\text{with} \qquad H(\boldsymbol{x}) = \sum_{j=1}^{N} h_j(x_j), \qquad \text{and} \qquad \mathcal{Q}(\boldsymbol{z}) = \sum_{i=1}^{M} q_i(z_j). \tag{30}$$

Here also $h_j(x_j)$ and $q_i(z_i)$ depend on the reference measures $\mu_x(\boldsymbol{x})$ and $\mu_n(\boldsymbol{x})$. The determination of α is not discussed.

3. Bayesian approach

3.1. Basics

The different steps of this approach are now well-known:

– From the observation model and the hypothesis (prior knowledge) on the noise derive the likelihood $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\beta})$;
– From the hypothesis (prior knowledge) on $\boldsymbol{x}$ derive the prior law $p(\boldsymbol{x}|\boldsymbol{\theta})$;
– Apply the Bayes rule to obtain $p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\beta},\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\beta})\,p(\boldsymbol{x}|\boldsymbol{\theta})/p(\boldsymbol{y};\boldsymbol{\beta},\boldsymbol{\theta})$;
– Define an estimation rule via a cost function $c(\boldsymbol{x},\widehat{\boldsymbol{x}})$ by:

$$\widehat{\boldsymbol{x}}(\boldsymbol{y};\boldsymbol{\beta},\boldsymbol{\theta}) = \arg\min_{\boldsymbol{z}} \left\{ \int c(\boldsymbol{x},\boldsymbol{z}) p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\beta},\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x} \right\}. \tag{31}$$

Different cost functions give different estimators:

– Maximum a posteriori (MAP):

$$C(\boldsymbol{x},\widehat{\boldsymbol{x}}) = 1 - \delta(\boldsymbol{x} - \widehat{\boldsymbol{x}}) \longrightarrow \widehat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \{p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta},\boldsymbol{\beta})\}. \tag{32}$$

– Posterior mean (PM):

$$C(\boldsymbol{x},\widehat{\boldsymbol{x}}) = [\boldsymbol{x} - \widehat{\boldsymbol{x}}]^t \boldsymbol{Q} [\boldsymbol{x} - \widehat{\boldsymbol{x}}]^t \longrightarrow \widehat{\boldsymbol{x}} = \mathrm{E}_{\boldsymbol{x}|\boldsymbol{y}}\{\boldsymbol{X}\} = \int \boldsymbol{x}\, p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta},\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{x}. \tag{33}$$

– Maximum of the Marginal a posteriori (MMAP):

$$C(\boldsymbol{x},\widehat{\boldsymbol{x}}) = \prod_j 1 - \delta(x_j - \widehat{x}_j) \longrightarrow \widehat{x}_j = \arg\max_{x_j} \{p(x_j|\boldsymbol{y};\boldsymbol{\theta})\}, \tag{34}$$

where

$$p(x_j|\boldsymbol{y};\boldsymbol{\theta}) = \int p(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{\theta})\,\mathrm{d}x_1 \cdots \mathrm{d}x_{j-1} \cdots \mathrm{d}x_{j+1} \cdots \mathrm{d}x_n. \tag{35}$$

To illustrate this, let consider the case of linear inverse problems $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}$ with the following hypothesis:

– $\boldsymbol{n}$ is zero-mean, white and Gaussian: $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, (1/\beta)\boldsymbol{I})$ which leads to:

$$p(\boldsymbol{y}|\boldsymbol{x}, \beta) \propto \exp\left[-\frac{1}{2}\beta\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2\right]. \tag{36}$$

– $\boldsymbol{x}$ is Gaussian: $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}_0, (1/\theta)\boldsymbol{P}_0)$:

$$p(\boldsymbol{x}|\theta) \propto \exp\left[-\frac{1}{2}\theta[\boldsymbol{x} - \boldsymbol{x}_0]^t \boldsymbol{P}_0^{-1}[\boldsymbol{x} - \boldsymbol{x}_0]\right]. \tag{37}$$

Then, using the Bayes rule it is easy to show that

$$\boldsymbol{x}|\boldsymbol{y} \sim \mathcal{N}(\widehat{\boldsymbol{x}}, \boldsymbol{P}) \quad \text{with} \quad \widehat{\boldsymbol{x}} = \boldsymbol{P}\boldsymbol{A}^t(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_0), \quad \boldsymbol{P} = \left(\boldsymbol{A}^t\boldsymbol{A} + \lambda\boldsymbol{P}_0^{-1}\right)^{-1}. \tag{38}$$

The MAP solution is:

$$\widehat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \{p(\boldsymbol{x}|\boldsymbol{y})\} = \arg\min_{\boldsymbol{x}} \{J(\boldsymbol{x})\}, \quad \text{with} \quad J(\boldsymbol{x}) = Q(\boldsymbol{x}) + \lambda\phi(\boldsymbol{x}), \tag{39}$$

where

$$Q(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2, \quad \phi(\boldsymbol{x}) = \boldsymbol{x}^t\boldsymbol{P}_0^{-1}\boldsymbol{x} = \|\boldsymbol{D}\boldsymbol{x}\|, \quad \lambda = \frac{\theta}{\beta} \tag{40}$$

Now, relaxing the second hypothesis, i.e; choosing other prior laws for $\boldsymbol{x}$ we obtain other MAP criteria. Let just note some special interesting cases:

– A Generalized Gaussian law for $\boldsymbol{x}$:

$$p(x_j) \propto \exp\left[-(x_j - m_j)^\alpha\right]. \tag{41}$$

The related MAP criterion becomes:

$$J(\boldsymbol{x}) = Q(\boldsymbol{x}) + \phi(\boldsymbol{x}) \quad \text{with} \quad \phi(\boldsymbol{x}) = \sum_j (x_j - m_j)^\alpha. \tag{42}$$

– A Gamma law for $\boldsymbol{x}$:

$$x_j \sim \mathcal{G}(\alpha, m_j) \longrightarrow p(x_j) \propto (x_j/m_j)^{-\alpha} \exp\left[-x_j/m_j\right]. \tag{43}$$

The related MAP criterion becomes:

$$J(\boldsymbol{x}) = Q(\boldsymbol{x}) + \phi(\boldsymbol{x}) \quad \text{with} \quad \phi(\boldsymbol{x}) = \sum_j \alpha \log\frac{x_j}{m_j} + \frac{x_j}{m_j}. \tag{44}$$

– A Beta law for $\boldsymbol{x}$:

$$x_j \sim \mathcal{B}(\alpha, \beta) \longrightarrow p(x_j) \propto x_j^\alpha (1 - x_j)^\beta. \tag{45}$$

The related MAP criterion becomes:

$$J(\boldsymbol{x}) = Q(\boldsymbol{x}) + \phi(\boldsymbol{x}) \quad \text{with} \quad \phi(\boldsymbol{x}) = \alpha\sum_j \log x_j + \beta\sum_j \log(1 - x_j). \tag{46}$$

– A Poisson law for $\boldsymbol{x}$:

$$p(x_j) \propto \frac{m_j^{x_j}}{x_j!} \exp\left[-m_j\right]. \tag{47}$$

The related MAP criterion becomes:

$$J(\boldsymbol{x}) = Q(\boldsymbol{x}) + \phi(\boldsymbol{x}) \quad \text{with} \quad \phi(\boldsymbol{x}) = -\sum_j x_j \log \frac{x_j}{m_j} + (x_j - m_j). \tag{48}$$

– Markovian models for $\boldsymbol{x}$:

$$J(\boldsymbol{x}) = Q(\boldsymbol{x}) + \phi(\boldsymbol{x}) \quad \text{with} \quad \phi(\boldsymbol{x}) = \alpha \sum_j \sum_{i \in N_j} V(x_j, x_i). \tag{49}$$

3.2. Extensions

The Bayesian approach can be exactly applied when all the direct (prior) probability laws ($p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\beta})$ and $p(\boldsymbol{x}|\boldsymbol{\theta})$) are assigned. Even, choosing an appropriate law is done in general by hand, another difficulty is to determine their parameters $(\boldsymbol{\beta}, \boldsymbol{\theta})$. This problem has been addressed by many authors and the subject is an active area in statistics. See [**?**, **?**, **?**, **?**], [**?**, **?**, **?**, **?**, **?**] and also [**?**, **?**, **?**].

All these methods can mainly be divided in three main families:

– Generalized MAP: In this approach one tries to estimate both the hyperparameters and the unknown variables $\boldsymbol{x}$ directly from the data by defining:

$$(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}}) = \arg \max_{(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\beta})} \{p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{y})\} \tag{50}$$

where

$$p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\beta})\, p(\boldsymbol{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, p(\boldsymbol{\beta}) \tag{51}$$

and where $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\beta})$ are appropriate prior laws. Many authors used the non informative prior law for them.

– Marginalization: In this approach one tries to estimate first the hyperparameters by marginalizing over the unknown variables $\boldsymbol{x}$:

$$p(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{y}) \propto p(\boldsymbol{\beta})\, p(\boldsymbol{\theta}) \int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\beta})\, p(\boldsymbol{x}|\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x} \tag{52}$$

and then, using them in the estimation of the unknown variables $\boldsymbol{x}$:

$$(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}}) = \arg \max_{(\boldsymbol{\theta}, \boldsymbol{\beta})} \{p(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{y})\} \tag{53}$$

– Nuisance parameters: In this approach the hyperparameters are considered as the nuisance parameters, so, marginalized:

$$p(\boldsymbol{x}|\boldsymbol{y}) = \int p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\beta})\, \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}\boldsymbol{\beta} \tag{54}$$

and the unknown variables $\boldsymbol{x}$ are estimated by:

$$\widehat{\boldsymbol{x}} = \arg \max_{\boldsymbol{x}} \{p(\boldsymbol{x}|\boldsymbol{y})\} \tag{55}$$

To see some more discussions and different possible implementations of these approaches see [**?**].

4. Discussed points in each approach

4.1. MEM:

– Choice of $\mathcal{C}$ and $\mu(\boldsymbol{x})$:
  – $\mathcal{C}$ must be a convex set, such as: $\mathbb{R}^N$, $\mathbb{R}^N_+$, $[a,b]^N$
  – Up to now, the whole analysis can be done for separable measures.
  – The only reference measures μ_j which permits to go through all the steps are those for which we have analytical expression for the Laplace transform of their logarithms.

– Accounting for the noise:
  In the first approach only the support and the energy of the noise is used. In the second approach we have more choices via the reference measures $\mu_n(\boldsymbol{n})$, but determination of α stays adhoc. In fact, in general, when the reference measures $\mu_n(\boldsymbol{n})$ and $\mu_x(\boldsymbol{x})$ depend on any parameters, this approach lacks any tool to determine them.

– Effective calculation of the solution:
  No problem, and more, this is probably the main interest of the approach which defines the solution, by construction, as the optimizer of a convex criterion.

– Characterization of the solution:
  A sensitivity analysis has been proposed by [**?**], but, in my opinion, this is not enough to characterize a solution to an inverse problem.
  It is not easy to use the notions of variance or covariance of the solution, because this approach does not define a posterior distribution for the solution.

– Possibility of the extension of the approach:
  I have not yet seen any extension of this approach to non linear inverse problems, or linear inverse problems in which the operator depend on unknown parameters, such as blind deconvolution or antenna array processing;
  The fact that we have to choose a convex set $\mathcal{C}$ on which the solution is defined excludes the inverse problems in which we know a priori that the solution is discrete-valued (binary, $n$ary images for example). This excludes the use of this approach in image segmentation or communication inverse problems (canal equalization, blind deconvolution, etc.).

4.2. BAYES:

– Choice of $p(\boldsymbol{x}|\boldsymbol{\beta})$:
  $p(\boldsymbol{x}|\boldsymbol{\beta})$ can be chosen separable or not. Evidently, separable $p(\boldsymbol{x}|\boldsymbol{\beta})$ (Entropic prior laws) simplifies the calculations. Accounting for correlations is easily done via Markovian models; In both cases (Entropic or Markovian prior laws), there are some tools for choosing them either by physical considerations, or by scale invariance arguments [**?**, **?**, **?**, **?**, **?**].

– Choice of the cost function or equivalently of an estimator MAP, PM, MMAP:
  This choice is done more on the basis of cost calculation. MAP calculation needs, in general, global optimization, but does not need any integration.
  MP or MMAP needs multidimensional integration, so in general, greater cost. However, there are approximate calculation techniques based on Monte Carlo

methods and Gibbs sampling.
- Effective calculation of different solutions:
  For MAP estimate, when the posterior law is unimodal, we can use any gradient descent based method, but if this is not the case, there are two categories of methods: Simulated Annealing or Deterministic relaxation (GNC). For more discussions on Bayesian calculations see [**?**] in this volume.

5. Comparisons and discussions

The following main items are discussed:

- In MEM, the unknowns $\boldsymbol{x}$ are considered as the mean values of a random vector $\boldsymbol{X}$ for which a prior probability measure $\mathrm{d}\mu(\boldsymbol{x})$ is defined.
- In BAYES, the unknowns $\boldsymbol{x}$ are considered as a sample of a random vector $\boldsymbol{X}$ for which a prior probability measure $p(\boldsymbol{x})$ is defined.
- In MEM, a probability distribution $p(\boldsymbol{x})$ is defined as the minimizer of the Kullback distance $K(p, \mu)$ subject to the data constraints, and the solution is defined to be $\mathrm{E}_p(\boldsymbol{X})$. What is interesting here is that this solution can equivalently be obtained as the minimizer of a convex criterion $J(\boldsymbol{x})$ subject to the data constraints, and what is more attractive is that, thanks to the convex analysis, this solution can also be obtained as the stationary point of a dual criterion which can easily be calculated numerically.
- In BAYES, the posterior law $p(\boldsymbol{x}|\boldsymbol{y})$ is calculated using the Bayes' rule. In fact, the data $\boldsymbol{y}$ are considered as a sample of a random vector $\boldsymbol{Y}$ for which we can define a conditional probability law $p(\boldsymbol{y}|\boldsymbol{x})$ which, when used in conjunction with the prior $p(\boldsymbol{x})$ in the Bayes' rule will give us the posterior law, from which we can define an estimator. One of these estimators is the posterior mean $\mathrm{E}_p(\boldsymbol{X})$, but others can also be defined. This posterior law is used not only to define an estimate (a solution), but also to calculate any other probabilistic information about that solution.
- In MEM, in its original version, the data are considered as the exact linear constraints. The uncertainty on the data are not considered, and consequently, the uncertainty on the solution is not handled. However, some extensions are recently presented to take account of the errors on the data and to calculate the sensitivity of the solution to these errors.
- In BAYES, the errors are naturally considered through $p(\boldsymbol{y}|\boldsymbol{x})$ and the uncertainty of the solution through the posterior probability $p(\boldsymbol{x}|\boldsymbol{y})$. Naturally then, we can compare the information content of the data and the prior model using their entropies. We can also measure the relative information content of the posterior to prior model by $K(p(\boldsymbol{x}|\boldsymbol{y}), p(\boldsymbol{x}))$.
- In MEM, even in their extended versions, it is not easy to handle with the hyperparameters. In BAYES, there are the necessary tools to handle them.
- In MEM, one can not yet handle with non linear problems. This is not the case of the BAYES.

As a final conclusion, we have to mention that, even if the two approaches are different, they can, in some cases result to the same definition of the solution as

the minimizer of the same criterion, and consequently, to give exactly the same numerical solutions to a given inverse problem. However, we can give different interpretations to the obtained result depending on the approach used to reach it. The main objective of this paper was to give a succinct presentation of the two approaches for the resolution of the inverse problems.

It is important to note that the two approaches give different views and interpretations which can be used advantageously for any application. Also, even the Bayesian approach is now really mature, the MEM approach is more recent. So, many of the conclusions I made today may be altered in future. In particular, new presentation of the method by Heinrich et al [**?**] in this volume will probably give new possibilities to the MEM method and will push the limits of the method to greatest generality.